dbGaP Study Release Notes

Release Notes for NHLBI TOPMed - NHGRI CCDG MGH AF, phs001062.v5.p2
"*NHLBI TOPMed - NHGRI CCDG: MGH Atrial Fibrillation Study*"

For any questions or comments, please contact: dbgap-help@ncbi.nlm.nih.gov.

October	20, 2016	Version 1 Data set release date
August	9, 2017	Version 2 Data set release date
May	29, 2018	Version 3 Data set release date
May	14, 2020	Version 4 Data set release date
May	28, 2021	Version 5 Data set release date

2021-05-28

Version 5 Data set release for NHLBI TOPMed - NHGRI CCDG MGH AF now available

This release includes the addition of Freeze 9 whole genome sequences (WGS) brokered through the Sequence Read Archive (SRA), and VCFs derived from WGS. Please refer to the latest study configuration report for a detailed description of each download component.

**There are no overlapping subjects between the 2 consent groups listed below.

Consent group 1 (c1): Health/Medical/Biomedical (IRB) (HMB-IRB)

Data Type	subjects	samples
Phenotype	908	907
Seq_DNA_SNP_CNV (VCFs)	908	907
WGS*	908	907

Consent group 2 (c2): Disease-Specific (Atrial Fibrillation, IRB, RD) (DS-AF-IRB-RD)

Data Type	subjects	samples
Phenotype	255	196
Seq_DNA_SNP_CNV (VCFs)	255	196
WGS*	255	196

*These data are brokered through the Sequence Read Archive (SRA). Please see Authorized Access instructions below.
For a description of non-SRA SAMPLE_USE terms, please see:
https://www.ncbi.nlm.nih.gov/projects/gap/submission/GetSampleUseTypes.cgi

Study and Phenotype Updates

1. **New Study Accession1**

   NHLBI TOPMed WGS MGH AF version 4 phs001062.v4.p2 has been updated to version 5. The dbGaP accession for the current phenotype data is **phs001062.v5.p2**. The participant number (p#) has not changed in version 5. No new subjects have been added to the study.

2. There are no updates to the phenotype datasets.

Molecular Data Updates

dbGaP QC steps for this release consist of checks for consistency of subject and sample IDs in phenotype and genotype components.

1. For samples and marker/enrichment-procedure info, see download components:
    a. phg001567.v1.TOPMed_WGS_MGH_AF_v5_frz9.sample-info.MULTI.tar.gz
2. Genotypes are available in a matrix format as multi-sample vcf file(s) packed within download component(s) marked as genotype-calls-vcf. Integrity of submitted vcf files and their compatibility with PSEQ are routinely checked. Components may be divided by platform and/or population.
    a. phg001567.v1.TOPMed_WGS_MGH_AF_v5_frz9.genotype-calls-vcf.WGS_markerset_grc38.c1.HMB-IRB.tar.gz
    b. phg001567.v1.TOPMed_WGS_MGH_AF_v5_frz9.genotype-calls-vcf.WGS_markerset_grc38.c2.DS-AF-IRB-RD.tar.gz

phg001400.v1	Freeze 8
phg001567.v1	Freeze 9

Authorized Access (Individual Level Data and SRA Data)

Individual level data and Sequence Read Archive (SRA) data are available for download through the dbGaP Authorized Access System upon approval of the Data Access Request (DAR):

- https://dbgap.ncbi.nlm.nih.gov/aa/wga.cgi?login=&page=login

Public FTP site (Summary Level Data Only)

All data tables, data dictionaries, and documents will be housed under one directory for ease of downloading. The data_dict filenames have an added study version number (phs#.v#) and deleted participant set number (p#) from the table accession (pht#.v#). The var_report filenames have an added study version number (phs#.v#). In the var_report files, variables contain version numbers (phv#.v#) and summaries were created for each consent group (c#). These FTP files are available at:

- https://ftp.ncbi.nlm.nih.gov/dbgap/studies/phs001062/phs001062.v5.p2

2020-05-14

Version 4 Data set release for NHLBI TOPMed - NHGRI CCDG MGH AF now available

This release includes the addition of Freeze 8 whole genome sequences (WGS) brokered through the Sequence Read Archive (SRA), VCFs derived from WGS. TOPMed and CCDG MGH AF have been combined into a single dbGaP study. Please refer to the latest study configuration report for a detailed description of each download component.

**There are no overlapping subjects between the 2 consent groups listed below.

Consent group 1 (c1): Health/Medical/Biomedical (IRB) (HMB-IRB)

Data Type	subjects	samples
Phenotype	908	907
Seq_DNA_SNP_CNV (VCFs)	908	907
WGS*	908	907

Consent group 2 (c2): Disease-Specific (Atrial Fibrillation, IRB, RD) (DS-AF-IRB-RD)

Data Type	subjects	samples
Phenotype	255	196
Seq_DNA_SNP_CNV (VCFs)	255	196
WGS*	255	196

*These data are brokered through the Sequence Read Archive (SRA). Please see Authorized Access instructions below.
For a description of non-SRA SAMPLE_USE terms, please see:
https://www.ncbi.nlm.nih.gov/projects/gap/submission/GetSampleUseTypes.cgi

Study and Phenotype Updates

1. **New Study Accession**

   NHLBI TOPMed WGS MGH AF version 3 phs001062.v3.p2 has been updated to version 4. The dbGaP accession for the current phenotype data is **phs001062.v4.p2**. The participant number (p#) has not changed in version 4. New subjects have been added to the study.

2. **Updated Datasets (n=3 datasets; all existing variables have been updated)**

pht	version	Dataset Name
5261	3	TOPMed_WGS_MGH_AF_Subject
5262	4	TOPMed_WGS_MGH_AF_Sample
5263	3	TOPMed_WGS_MGH_AF_Sample_Attributes

Molecular Data Updates

dbGaP QC steps for this release consist of checks for consistency of subject and sample IDs in phenotype and genotype components.

1. For samples and marker/enrichment-procedure info, see download components:
   a. phg001400.v1.TOPMed_WGS_MGH_AF_v4.sample-info.MULTI.tar.gz
   b. phg001400.v1.TOPMed_WGS_MGH_AF_v4.marker-info.MULTI.tar.gz
2. Genotypes are available in a matrix format as multi-sample vcf file(s) packed within download component(s) marked as genotype-calls-vcf. Integrity of submitted vcf files and their compatibility with PSEQ are routinely checked. Components may be divided by platform and/or population.
   a. phg001400.v1.TOPMed_WGS_MGH_AF_v4.genotype-calls-vcf.WGS_markerset_grc38.c2.DS-AF-IRB-RD.tar.gz
   b. phg001400.v1.TOPMed_WGS_MGH_AF_v4.genotype-calls-vcf.WGS_markerset_grc38.c1.HMB-IRB.tar.gz.
3. Only Freeze 5b and Freeze 8 VCFs will be available for download.

Authorized Access (Individual Level Data and SRA Data)

Individual level data and Sequence Read Archive (SRA) data are available for download through the dbGaP Authorized Access System upon approval of the Data Access Request (DAR):

- https://dbgap.ncbi.nlm.nih.gov/aa/wga.cgi?login=&page=login

Public FTP site (Summary Level Data Only)

All data tables, data dictionaries, and documents will be housed under one directory for ease of downloading. The data_dict filenames have an added study version number (phs#.v#) and deleted participant set number (p#) from the table accession (pht#.v#). The var_report filenames have an added study version number (phs#.v#). In the var_report files, variables contain version numbers (phv#.v#) and summaries were created for each consent group (c#). These FTP files are available at:

- https://ftp.ncbi.nlm.nih.gov/dbgap/studies/phs001062/phs001062.v4.p2

2018-05-29
Version 3 Data set release for NHLBI TOPMed WGS MGH AF now available

This release includes a second genotype call set (GRCh38). Please refer to the latest study configuration report for a detailed description of each download component.

**There are no overlapping subjects between the 2 consent groups listed below.

Consent group 1 (c1): Health/Medical/Biomedical (IRB) (HMB-IRB)

	Phenotype	Seq_DNA_SNP_CNV (VCFs)	Seq_DNA_WholeGenome
subjects	908	834	722 and ongoing
samples	908	834	722 and ongoing

Consent group 2 (c2): Disease-Specific (Atrial Fibrillation, IRB, RD) (DS-AF-IRB-RD)

	Phenotype	Seq_DNA_SNP_CNV (VCFs)	Seq_DNA_WholeGenome
subjects	91	84	71 and ongoing
samples	91	84	71 and ongoing

Molecular data descriptions:
(https://www.ncbi.nlm.nih.gov/projects/gap/submission/GetSampleUseTypes.cgi)

a. Seq_DNA_WholeGenome: Whole genome sequencing
b. Seq_DNA_SNP_CNV: SNP and CNV genotypes derived from sequence data (VCFs)

Study and Phenotype Updates

1. **New Study Accession**

   NHLBI TOPMed WGS MGH AF version 2 phs001062.v2.p2 has been updated to version 3. The dbGaP accession for the current phenotype data is **phs001062.v3.p2**. The participant number (p#) has not changed in version 3. There are new subjects added to the study.

2. There are no changes to the phenotype data since the last version release. Please note we are discontinuing the submission and distribution of the SAMPLE_USE variable. The sample use counts will be populated by SRA (sequences) and dbGaP (all other submitted molecular data).

Molecular Data Updates

dbGaP QC steps for this release consist of checks for consistency of subject and sample IDs in phenotype and genotype components.

1. For samples and marker/enrichment-procedure info, see download components:

   a. phg001057.v1.TOPMed_WGS_MGH_AF_v3.sample-info.MULTI.tar.gz
   b. phg001057.v1.TOPMed_WGS_MGH_AF_v3.marker-info.MULTI.tar.gz

2. The standard dbGaP QC pipeline was applied on SNP genotypes in PLINK format. Results are in tar files marked as "genotype-qc"
3. Genotypes are available in a matrix format as multi-sample .vcf file(s) packed within download component(s) marked as "genotype-calls-vcf. Integrity of submitted .vcf files and their compatibility with PSEQ are routinely checked. Components may be divided by platform and/or population.
   a. phg001057.v1.TOPMed_WGS_MGH_AF_v3.genotype-calls-vcf.WGS_markerset_grc38.c1.HMB-IRB.tar.gz
   b. phg001057.v1.TOPMed_WGS_MGH_AF_v3.genotype-calls-vcf.WGS_markerset_grc38.c2.DS-AF-IRB-RD.tar.gz.

Authorized Access (Individual Level Data and SRA Data)

Individual level data and Sequence Read Archive (SRA) data are available for download through the dbGaP Authorized Access System upon approval of the Data Access Request (DAR):

- https://dbgap.ncbi.nlm.nih.gov/aa/wga.cgi?login=&page=login

Public FTP site (Summary Level Data Only)

All data tables, data dictionaries, and documents will be housed under one directory for ease of downloading. The data_dict filenames have an added study version number (phs#.v#) and deleted participant set number (p#) from the table accession (pht#.v#). The var_report filenames have an added study version number (phs#.v#). In the var_report files, variables contain version numbers (phv#.v#) and summaries were created for each consent group (c#). These FTP files are available at:

- https://ftp.ncbi.nlm.nih.gov/dbgap/studies/phs001062/phs001062.v3.p2

2017-08-09

Version 2 Data set release for NHLBI TOPMed WGS MGH AF now available

This release includes TOPMed Phase I phenotype tables, whole genome sequences (WGS) brokered through the SRA, and VCFs derived from WGS. Additionally, phenotype tables include subjects and samples beyond TOPMed Phase I in order to instantiate IDs for future versions. Please refer to the latest study configuration report for a detailed description of each download component.

Consent group 1 (c1): Health/Medical/Biomedical (IRB) (HMB-IRB)

	phenotype	SRA	VCFs
subjects	908	722	721
samples	908	722	721

Consent group 2 (c2): Disease-Specific (Atrial Fibrillation, IRB, RD) (DS-AF-IRB-RD)

	phenotype	SRA	VCFs
subjects	91	71	71

samples	91	71	71

Study and Phenotype Updates

1. **New Study Accession**

   NHLBI TOPMed WGS MGH AF version 1 phs001062.v1.p1 has been updated to version 2. The dbGaP accession for the current phenotype data is **phs001062.v2.p2**. The participant number (p#) has changed in version 2; subjects have been retired and/or changed between consent groups. There are new subjects added to the study.

2. **Updated Datasets (n=3)**
   a. pht005261.v2.p2 TOPMed_WGS_MGH_AF_Subject
   b. pht005262.v2.p2 TOPMed_WGS_MGH_AF_Sample
   c. pht005263.v2.p2 TOPMed_WGS_MGH_AF_Sample_Attributes

3. **Retired Datasets (n=1)**
   a. pht005689.v1.p1 TOPMed_WGS_MGH_AF_Subject_Phenotypes*
   *Subject phenotype and pedigree data will be available through the MGH Cohort study (phs001001) in the near future.

4. **New Variables (n=7)**

pht	pht version	Dataset Name	phv	Variable Name
5261	2	TOPMed_WGS_MGH_AF_Subject	309962	SUBJECT_SOURCE
5261	2	TOPMed_WGS_MGH_AF_Subject	309963	SOURCE_SUBJECT_ID
5263	2	TOPMed_WGS_MGH_AF_Sample_Attributes	309964	SEQUENCING_CENTER
5263	2	TOPMed_WGS_MGH_AF_Sample_Attributes	309965	Funding_Source
5263	2	TOPMed_WGS_MGH_AF_Sample_Attributes	309966	TOPMed_Project
5263	2	TOPMed_WGS_MGH_AF_Sample_Attributes	309967	Study_Name
5263	2	TOPMed_WGS_MGH_AF_Sample_Attributes	309968	TOPMed_Phase

Molecular Data Updates

dbGaP QC steps for this release consist of checks for consistency of subject and sample IDs in phenotype and genotype components.

1. For samples and marker/enrichment-procedure info, see download components:
   a. phg000924.v1.TOPMed_WGS_MGH_AF_v2.sample-info.MULTI.tar.gz
   b. phg000924.v1.TOPMed_WGS_MGH_AF_v2.marker-info.MULTI.tar.gz
2. The standard dbGaP QC pipeline was applied on SNP genotypes in PLINK format. Results are in tar files marked as "genotype-qc"
3. Genotypes are available in a matrix format as multi-sample .vcf file(s) packed within download component(s) marked as "genotype-calls-vcf. Integrity of submitted .vcf files and their compatibility with PSEQ are routinely checked. Components may be divided by platform and/or population.
   a. phg000924.v1.TOPMed_WGS_MGH_AF_v2.genotype-calls-vcf.WGS_markerset_grc37.c2.DS-AF-IRB-RD.tar.gz
   b. phg000924.v1.TOPMed_WGS_MGH_AF_v2.genotype-calls-vcf.WGS_markerset_grc37.c1.HMB-IRB.tar.gz.

Authorized Access (Individual Level Data and SRA Data)

Individual level data and Sequence Read Archive (SRA) data are available for download through the dbGaP Authorized Access System upon approval of the Data Access Request (DAR):

- https://dbgap.ncbi.nlm.nih.gov/aa/wga.cgi?login=&page=login

Public FTP site (Summary Level Data Only)

All data tables, data dictionaries, and documents will be housed under one directory for ease of downloading. The data_dict filenames have an added study version number (phs#.v#) and deleted participant set number (p#) from the table accession (pht#.v#). The var_report filenames have an added study version number (phs#.v#). In the var_report files, variables contain version numbers (phv#.v#) and summaries were created for each consent group (c#). These FTP files are available at:

- https://ftp.ncbi.nlm.nih.gov/dbgap/studies/phs001062/phs001062.v2.p2

2016-10-20

Version 1 Data set release for NHLBI TOPMed WGS MGH AF now available

This release includes TOPMed Phase I phenotype tables, whole genome sequences (WGS) brokered through the SRA, and VCFs derived from WGS. Additionally, phenotype tables include subjects and samples beyond TOPMed Phase I in order to instantiate IDs for future versions. Please refer to the latest study configuration report for a detailed description of each download component.

Consent group 1 (c1): Health/Medical/Biomedical (IRB) (HMB-IRB)

	phenotype	SRA/VCFs
subjects	723	243
samples	723	243

Consent group 2 (c2): Disease-Specific (Atrial Fibrillation, IRB, RD) (DS-AF-IRB-RD)

	phenotype	SRA/VCFs
subjects	71	31
samples	71	31

Molecular Data Updates

dbGaP QC steps for this release consisted of checks for consistency of subject and sample IDs in phenotype and genotype components:

1. For samples and marker/enrichment-procedure info see download components:
   a. phg000805.v1.TOPMed_WGS_MGH_AF.sample-info.MULTI.tar.gz
   b. phg000805.v1.TOPMed_WGS_MGH_AF.marker-info.MULTI.tar.gz
2. Genotypes are available in a matrix format as multi-sample .vcf file(s) packed within download component(s) marked as "genotype-calls-vcf. Integrity of submitted .vcf files and

their compatibility with PSEQ are routinely checked. It is noted when components are divided by platform and/or population.

   a. phg000805.v1.TOPMed_WGS_MGH_AF.genotype-calls-vcf.WGS_markerset_grc37.c2.DS-AF-IRB-RD.tar.gz
   b. phg000805.v1.TOPMed_WGS_MGH_AF.genotype-calls-vcf.WGS_markerset_grc37.c1.HMB-IRB.tar.gz

3. The standard dbGaP QC pipeline was applied on SNP genotypes in PLINK format. Results are in tar files marked "genotype-qc".

Authorized Access (Individual Level Data and SRA Data)

Individual level data and SRA sequencing data are available for download through the dbGaP Authorized Access System upon approval of the Data Access Request (DAR):

- https://dbgap.ncbi.nlm.nih.gov/aa/wga.cgi?login=&page=login

Public FTP site (Summary Level Data Only)

All data tables, data dictionaries, and documents will be housed under one directory for ease of downloading. The data_dict filenames have an added study version number (phs#.v#) and deleted participant set number (p#) from the table accession (pht#.v#). The var_report filenames have an added study version number (phs#.v#). In the var_report files, variables contain version numbers (phv#.v#) and summaries were created for each consent group (c#). These FTP files are available at:

- https://ftp.ncbi.nlm.nih.gov/dbgap/studies/phs001062/phs001062.v1.p1